

Analytic approach to variance optimization under an ℓ_1 constraint

Imre Kondor^{1,2,3}, Gábor Papp⁴, and Fabio Caccioli^{2,5,6,a}

¹ Parmenides Foundation, Pullach, Germany

² London Mathematical Laboratory, London, UK

³ Complexity Science Hub, Vienna, Austria

⁴ Eötvös Loránd University, Institute for Physics, Budapest, Hungary

⁵ University College London, Department of Computer Science, London, WC1E 6BT, UK

⁶ Systemic Risk Centre, London School of Economics and Political Sciences, London, UK

Received 13 July 2018 / Received in final form 17 October 2018

Published online 14 January 2019

© The Author(s) 2019. This article is published with open access at [Springerlink.com](https://www.springerlink.com)

Abstract. The optimization of the variance of a portfolio of N independent but not identically distributed assets, supplemented by a budget constraint and an asymmetric ℓ_1 regularizer, is carried out analytically by the replica method borrowed from the theory of disordered systems. The asymmetric regularizer allows us to penalize short and long positions differently, so the present treatment includes the no-short-constrained portfolio optimization problem as a special case. Results are presented for the out-of-sample and the in-sample estimator of the regularized variance, the relative estimation error, the density of the assets eliminated from the portfolio by the regularizer, and the distribution of the optimal portfolio weights. We have studied the dependence of these quantities on the ratio r of the portfolio's dimension N to the sample size T , and on the strength of the regularizer. We have checked the analytic results by numerical simulations, and found general agreement. Regularization extends the interval where the optimization can be carried out, and suppresses the large sample fluctuations, but the performance of ℓ_1 regularization is rather disappointing: if the sample size is large relative to the dimension, i.e. r is small, the regularizer does not play any role, while for r 's where the regularizer starts to be felt the estimation error is already so large as to make the whole optimization exercise pointless. We find that the ℓ_1 regularization can eliminate at most half the assets from the portfolio (by setting their weights to exactly zero), corresponding to this there is a critical ratio $r = 2$ beyond which the ℓ_1 regularized variance cannot be optimized: the regularized variance becomes constant over the simplex. These facts do not seem to have been noticed in the literature.

1 Introduction

In this paper, we present analytic results for a simple quadratic optimization problem with a linear constraint plus an ℓ_1 regularizer. Although we are going to speak in terms of portfolio optimization, it is important to emphasize that the problem we address is not specific to portfolios, but is a generic feature of quadratic optimization if the dimension is high and the objective function is estimated on the basis of a limited number of observations. We will assume that there is no additional information (like prior knowledge or sparsity) available besides the observations and wish to find out how much can be learned from the limited data. Our objective function will be the portfolio variance. In order to find the optimum of the variance over the portfolio weights, one has to invert the estimated covariance matrix, which is possible only if its dimension N is not larger than the number of observations T . The ratio $r = N/T$ is a fundamentally important control parameter of the problem. If the number of observations is much larger than the dimension,

classical statistics works and the estimated optimum will be very close to the true optimum which can be obtained in the limit $T \rightarrow \infty$. If T is not very large relative to N , we are in the high-dimensional regime where sample fluctuations can be large and regularizers have to be introduced to rein them in. Regularizers suppress large excursions, and unavoidably introduce bias, but the hope is that a reasonable trade-off can be achieved between the bias and sample fluctuations with a proper choice of the strength of the regularizer. To see whether this hope is fulfilled is one of the aims of this paper.

A common regularizer is ℓ_2 (shrinkage or ridge regression) whose effect has been studied by a number of authors, see [1–7] among many others. In its most recent nonlinear form shrinkage can produce very good quality estimates [8–10]. Another popular regularizer is based on the ℓ_1 norm (lasso) [11]. Lasso is known to lead to sparse estimates, reducing the effective dimension of the problem and stabilizing the estimator. Jagannathan and Ma [12] considered portfolio optimization under a constraint excluding short positions. Although they did not speak about regularization, a no-short constraint is, in fact, a special case of an asymmetric ℓ_1 regularizer.

^a e-mail: f.caccioli@ucl.ac.uk

Brodie et al. [13] and DeMiguel et al. [14] studied the effect of ℓ_1 regularization on the performance and stability of portfolio selection. Subsequently, a number of groups investigated various aspects of the application of ℓ_1 and related regularizers in portfolio optimization, e.g. [15–19].

The problem of optimizing the variance under an ℓ_1 constraint is a quadratic programming task which can be solved numerically. Our purpose here is to solve this problem analytically, which, to the best of our knowledge, has not been done before. The method that enables us to do this is borrowed from the theory of disordered systems and goes by the name of the method of replicas [20]. It assumes that the underlying distribution is Gaussian and it works in the Kolmogorov limit, where both N and T go to infinity, but their ratio $r = N/T$ is kept finite.

We will show that the ℓ_1 regularizer does not eliminate the instability, only shifts its value. (A similar effect was observed in the case of the Expected Shortfall risk measure in [21].) The new critical value turns out to be $r = 2$, corresponding to the fact that ℓ_1 sets to strictly zero at most half the portfolio weights. (Of course, in a strongly heterogeneous portfolio the high volatility elements can be suppressed, with their weights rendered very small in the optimal portfolio. We distinguish this from the elimination of some of the dimensions by ℓ_1 , with the corresponding weights becoming exactly zero, even if there is little or no difference between their true volatilities.)

There is an important difference between our analytic approach and the standard statistical estimation procedure which analyzes a given sample and tests it by cross validation [22]. Instead, our method allows us to average over the whole ensemble of samples. Corresponding to this, the step-like effect of ℓ_1 , eliminating the dimensions one by one, is replaced upon averaging over the samples by a smooth, monotonically increasing density of the zero weights.

In order to make contact with a previous work in which we treated the case of excluded short positions [23], we are going to consider an asymmetric ℓ_1 regularizer here, with different slopes for positive, resp. negative weights. We find that in the most important results only the right hand side slope appears.

Many dimensionality reduction or cleaning methods focus on the covariance matrix, especially on its spectrum. In contrast, the special version of the replica method we use allows us to derive the distribution of optimal portfolio weights directly.

The plan of the paper is as follows. In Section 2, we set up the problem and present some preliminary results. In Section 3, we recall some results from [23] where the task of optimizing over the N portfolio weights has been reduced to the optimization of an effective objective function depending on five order parameters. We also spell out the first order conditions (or saddle point conditions) that determine the stationary point of the objective function. The solution to the saddle point equations is analyzed in a number of subsections and the results for various special cases are displayed graphically. Section 4 is a summary of the results, while a sketch of the derivation of the effective objective function is provided in the Appendix.

2 Preliminaries

In this section, we set up the optimization problem, fix notation and present some preliminary results that will be useful as checks on the replica calculation later.

We consider a portfolio of N assets with random returns x_i , $i = 1, \dots, N$. For simplicity, we assume that the returns are independent Gaussian random variables with zero expectation value and variance σ_i^2 , which may be different for each asset i . For the time being we assume that we have complete knowledge of the distribution of the returns. If we denote the portfolio weights as w_i , the return on the portfolio is $\sum_{i=1}^N w_i x_i$, and under the assumption above the variance of the portfolio will be

$$\sigma_p^2 = \sum_i \sigma_i^2 w_i^2. \quad (1)$$

This is to be minimized subject to the budget constraint

$$\sum_i w_i = N, \quad (2)$$

where we set the budget to be N instead of the usual 1, to have $\mathcal{O}(1)$ weights in the limit of large N .

Then, with the Lagrange multiplier associated with the budget constraint denoted by λ we would have to find the minimum of

$$F = \sum_i \sigma_i^2 w_i^2 - \lambda \left(\sum_{i=1}^N w_i - N \right) \quad (3)$$

over the weights w_i , a trivial task.

So far, the distribution of the returns (in particular, the variances of the assets σ_i^2) have been assumed to be known. In real life this is never the case, instead we have to estimate the optimal weights and portfolio variance on the basis of finite samples. Let us assume that we draw these samples from a multivariate distribution of independent Gaussian variables with individual standard deviations σ_i . These samples are constituted of T observations for each asset: x_{it} , $i = 1, 2, \dots, N$; $t = 1, 2, \dots, T$. We wish to learn to what extent it is possible to recover the true optimum of the variance and the optimal weights by averaging over a large number of such samples.

Thus, we have the optimization problem

$$\min_{w_i} \left(\sum_{i,j} w_i C_{ij} w_j \right), \text{ s.t. } \sum_{i=1}^N w_i = N, \quad (4)$$

where C_{ij} is the estimated covariance matrix

$$C_{ij} = \frac{1}{T} \sum_{t=1}^T x_{it} x_{jt}. \quad (5)$$

Substituting (5) into (4) the optimization problem becomes

$$\min_{w_i} \left(\frac{1}{T} \sum_{t=1}^T \left(\sum_i w_i x_{it} \right)^2 \right), \text{ s.t. } \sum_{i=1}^N w_i = N. \quad (6)$$

This is a quadratic optimization problem which can be solved numerically, as long as the covariance matrix is positive definite, which holds with probability one for $T \geq N$, that is for $r < 1$.

As r approaches 1 from below, sample fluctuations become larger and larger, until at $r = 1$ the estimation error diverges, and for $r > 1$ the optimization becomes meaningless. In order to tame the large sample fluctuations, it is a standard procedure to introduce regularizers that suppress large excursions of the estimated weights (at the price of introducing bias).

The regularizer we wish to use here is based on the ℓ_1 norm (lasso) [11]. It is known to result in sparse estimates, which in the present context means eliminating a part of the assets from the optimal portfolio, thereby reducing its effective dimension. Lasso is extensively used in a variety of problems in high-dimensional statistics and machine learning [22,24]. Its first applications to portfolio optimization is due to Brodie et al. [13] and DeMiguel et al. [14]. For the non-analytic character of lasso a full analytic treatment has, to our knowledge, not been attempted. An analytic approach valid in the large N limit will be presented in the next section.

Let us spell out the ℓ_1 regularizer we are going to apply:

$$\ell_1(\eta_1, \eta_2) = \eta_1 \sum_i w_i \theta(w_i) - \eta_2 \sum_i w_i \theta(-w_i), \quad (7)$$

where η_1 and η_2 are positive coefficients and $\theta(x)$ is the Heaviside function. The regularizer so defined is asymmetric, having different slopes for positive and negative weights. The special case $\eta_1 = \eta_2 = \eta$ corresponds to the usual expression $\eta \sum_i |w_i|$. Keeping the two slopes different allows us to penalize long and short positions differently.

Our regularized objective function is then

$$F = \frac{1}{T} \sum_{t=1}^T \left(\sum_i w_i x_{it} \right)^2 + \eta_1 \sum_i w_i \theta(w_i) - \eta_2 \sum_i w_i \theta(-w_i) - \lambda \left(\sum_{i=1}^N w_i - N \right). \quad (8)$$

As the first term is non-negative and the last term vanishes for w_i 's satisfying the budget constraint, F is larger or equal to the minimum of $\ell_1(\eta_1, \eta_2)$, which is $N\eta_1$. Therefore $F \geq N\eta_1$, where the equality holds when the variance vanishes and the weights minimize the regularizer $\ell_1(\eta_1, \eta_2)$ (which requires that they are on the simplex $w_i \geq 0, \forall i, \sum_i w_i = N$). Alternatively, for the value of the

objective function per asset we have the inequality

$$\frac{F}{N} = f \geq \eta_1. \quad (9)$$

This will prove important later.

As it stands, (8) is amenable for numerical work, with the returns drawn from a suitable distribution. When the returns are independent Gaussians and N and T are large, one can derive the analytic results displayed in the next section.

A special limit of the above optimization problem is already worth considering at this point, because it provides an important consistency check on the results to be presented later: Let us assume that we have very large samples compared with the number of assets in the portfolio, i.e. $T \gg N$, or $r = N/T \rightarrow 0$. This means we have complete information about the distribution of returns. Then, for the independent random returns considered here, the covariance matrix C_{ij} becomes diagonal with diagonal elements σ_i^2 , and the optimization problem becomes

$$F = \sum_i \sigma_i^2 w_i^2 + \eta_1 \sum_i w_i \theta(w_i) - \eta_2 \sum_i w_i \theta(-w_i) - \lambda \left(\sum_{i=1}^N w_i - N \right). \quad (10)$$

A little reflection shows that the solution of this optimization problem can satisfy the budget constraint $\sum_i w_i = N$ for a positive N only if the Lagrange multiplier λ is larger than the *right* slope η_1 of the regularizer: $\lambda > \eta_1$. Then the Lagrange multiplier works out to be

$$\lambda = \frac{2N}{\sum_{j=1}^N 1/\sigma_j^2} + \eta_1, \quad (11)$$

the optimal weights

$$w_i^* = \frac{N}{\sum_{j=1}^N 1/\sigma_j^2} \frac{1}{\sigma_i^2}, \quad (12)$$

and the minimal value of the objective function F obtains as

$$F^* = \frac{N^2}{\sum_{j=1}^N 1/\sigma_j^2} + N\eta_1, \quad (13)$$

while the minimal value of the objective function per asset is

$$\frac{F^*}{N} = f^* = \frac{N}{\sum_{j=1}^N 1/\sigma_j^2} + \eta_1 = \frac{\lambda + \eta_1}{2}. \quad (14)$$

Note the order of magnitudes in the above formulae: λ , $\eta_{1,2}$ and w_i^* are of $\mathcal{O}(1)$, the sum $\sum_{j=1}^N 1/\sigma_j^2$ and the objective function are $\mathcal{O}(N)$. (Here and in most of the following, we are assuming that the standard deviations

σ_j^2 are of order one. When one of the assets is riskless, the corresponding σ_j^2 goes to zero, and one has to pay special attention to this exceptional asset.) We also have to point out that there is a difference in the notation relative to our earlier papers, especially [23], where we absorbed a factor $1/2r$ in the definition of the objective function f^* and the Lagrange multiplier λ . This did not change any of the results there, except sending λ to infinity in the limit $r \rightarrow 0$, which resulted in some convenience. In contrast to that paper, instead of considering the special limit $\eta_1 = 0$ and $\eta_2 \rightarrow \infty$, here we are going to keep the coefficients of the regularizer finite, so the convention of absorbing $1/2r$ into the objective function would dictate its absorption into η_1, η_2 as well. This would distort some of the figures, and would make the message of the paper harder to grasp. Therefore, in the present paper we have this factor $1/2r$ explicitly written out and kept throughout the paper.

The results obtained above for the Lagrange multiplier, the optimal weights and the optimal value of the objective function in the limit $r \rightarrow 0$ are the true values for these quantities that would be obtained over an infinitely long observation time, when sample fluctuations become irrelevant. Likewise, in the same limit the distribution $p(w)$ of the optimal portfolio weights would be a series of sharp spikes

$$p(w) = \frac{1}{N} \sum_{i=1}^N \delta(w - w_i^*), \quad (15)$$

where $\delta(x)$ is the Dirac delta distribution.

3 Results for the variance optimized under an ℓ_1 constraint

Our task is to find the optimum of the objective function in (8), where the returns x_{it} are assumed to be drawn from the joint probability density of N independent Gaussian variables with zero mean and variance σ_i^2 . Following the special version of the replica method laid out in [21], in [23] we showed how the optimization of (8) could be reduced to that of an effective objective function depending on five “order parameters”. The method we applied to achieve this was the method of replicas, borrowed from the statistical mechanics of disordered systems [20]. (We will denote this effective objective function by the same symbol f as its full-information counterpart in the preceding section, and will omit the adjective “effective” in the following.)

The derivation has been presented in [21] and also in the appendices of [23], and is sketched in the Appendix to this paper for easier reference. In the present section, we can start from the expression for the effective objective function $f(\lambda, q_0, \Delta, \hat{q}_0, \hat{\Delta})$ depending on the order parameters $\lambda, q_0, \Delta, \hat{q}_0, \hat{\Delta}$, as given in the Appendix:

$$f(\lambda, q_0, \Delta, \hat{q}_0, \hat{\Delta}) = \frac{q_0}{(1 + \Delta)} - 2r\hat{q}_0\Delta - 2r\hat{\Delta}q_0 + \lambda + \min_{\vec{w}} \langle V(\vec{w}) \rangle_{z, \sigma}, \quad (16)$$

where

$$V = 2r\hat{\Delta}\sigma^2w^2 - 2rwz\sigma\sqrt{-2\hat{q}_0} - \lambda w + \eta_1w\theta(w) - \eta_2w\theta(-w) \quad (17)$$

and the double average $\langle \dots \rangle_{z, \sigma}$ means

$$\int_0^\infty d\sigma \frac{1}{N} \sum_i \delta(\sigma - \sigma_i) \int_{-\infty}^\infty \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} \dots, \quad (18)$$

and σ_i is the standard deviation of the distribution of returns on asset i .

The minimum of the “potential” V is at

$$w^* = \frac{2r\sigma z\sqrt{-2\hat{q}_0} + \lambda - \eta_1\theta(w^*) + \eta_2\theta(-w^*)}{4r\hat{\Delta}\sigma^2}. \quad (19)$$

Substituting this back into (17) and performing the averaging according to the recipe (18) we find

$$\min_{\vec{w}} \langle V(\vec{w}) \rangle_{z, \sigma} = \frac{2r\hat{q}_0}{\hat{\Delta}} \frac{1}{N} \sum_i \left(W \left(\frac{\lambda - \eta_1}{2r\sigma_i\sqrt{-2\hat{q}_0}} \right) + W \left(-\frac{\lambda + \eta_2}{2r\sigma_i\sqrt{-2\hat{q}_0}} \right) \right). \quad (20)$$

This is then the explicit form of the last term in (16), which thus becomes

$$f(\lambda, q_0, \Delta, \hat{q}_0, \hat{\Delta}) = \frac{q_0}{(1 + \Delta)} - 2r\hat{q}_0\Delta - 2r\hat{\Delta}q_0 + \lambda + \frac{2r\hat{q}_0}{\hat{\Delta}} \frac{1}{N} \sum_i \left(W \left(\frac{\lambda - \eta_1}{2r\sigma_i\sqrt{-2\hat{q}_0}} \right) + W \left(-\frac{\lambda + \eta_2}{2r\sigma_i\sqrt{-2\hat{q}_0}} \right) \right). \quad (21)$$

The function W appearing here is the third integral of the standard normal Gaussian density; its precise definition will be given shortly, together with two more functions that appear frequently in the following.

Stationarity of (21) with respect to the order parameters gives the first order conditions

$$\hat{\Delta} = \frac{1}{2r(1 + \Delta)} \quad (22)$$

$$\hat{q}_0 = -\frac{q_0}{2r(1 + \Delta)^2} \quad (23)$$

$$\frac{1}{\sqrt{q_0 r}} = \frac{1}{N} \sum_i \frac{1}{\sigma_i} \left(\Psi \left(\frac{w_1^{(i)}}{\sigma_w^{(i)}} \right) - \Psi \left(-\frac{w_2^{(i)}}{\sigma_w^{(i)}} \right) \right) \quad (24)$$

$$\Delta = \frac{\frac{r}{N} \sum_i \left(\Phi \left(\frac{w_1^{(i)}}{\sigma_w^{(i)}} \right) + \Phi \left(-\frac{w_2^{(i)}}{\sigma_w^{(i)}} \right) \right)}{1 - \frac{r}{N} \sum_i \left(\Phi \left(\frac{w_1^{(i)}}{\sigma_w^{(i)}} \right) + \Phi \left(-\frac{w_2^{(i)}}{\sigma_w^{(i)}} \right) \right)} \quad (25)$$

$$\frac{1}{2r} = \frac{1}{N} \sum_i \left(W \left(\frac{w_1^{(i)}}{\sigma_w^{(i)}} \right) + W \left(-\frac{w_2^{(i)}}{\sigma_w^{(i)}} \right) \right). \quad (26)$$

Here $r = N/T$, as before. The functions Φ , Ψ and W are the integrals of the Gaussian density:

$$\Phi(x) = \int_{-\infty}^x \frac{dt}{\sqrt{2\pi}} e^{-t^2/2}, \quad (27)$$

$$\Psi(x) = \int_{-\infty}^x dt \Phi(t), \quad (28)$$

$$W(x) = \int_{-\infty}^x dt \Psi(t). \quad (29)$$

In the above formulae the following notations have been introduced:

$$w_1^{(i)} = \frac{\lambda - \eta_1}{4r\sigma_i^2\hat{\Delta}} = \frac{(\lambda - \eta_1)(1 + \Delta)}{2\sigma_i^2}, \quad (30)$$

$$w_2^{(i)} = \frac{\lambda + \eta_2}{4r\sigma_i^2\hat{\Delta}} = \frac{(\lambda + \eta_2)(1 + \Delta)}{2\sigma_i^2}, \quad (31)$$

and

$$\sigma_w^{(i)} = \frac{\sqrt{q_0 r}}{\sigma_i}, \quad (32)$$

where in (30), (31) and (32) use has been made of (22) and (23). With this we can eliminate \hat{q}_0 and $\hat{\Delta}$ from our equations. Proceeding similarly in (21) and using (26) we find the expression for the objective function in terms of the remaining three order parameters as

$$f = \lambda - \frac{q_0}{2r(1 + \Delta)^2}. \quad (33)$$

According to the derivation of the objective function in the Appendix, when (24), (25) and (26) are solved and λ , q_0 and Δ are obtained as functions of the control parameters r , η_1 and η_2 , equation (33) gives the *in-sample estimate* of the objective function.

With (19) the distribution of weights obtains from $p(w) = \langle \delta(w - w^*) \rangle_{z\sigma}$ as

$$\begin{aligned} p(w) = n_0 \delta(w) + \frac{1}{N} \sum_i \frac{1}{\sigma_w^{(i)} \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{w - w_1^{(i)}}{\sigma_w^{(i)}} \right)^2} \theta(w) \\ + \frac{1}{N} \sum_i \frac{1}{\sigma_w^{(i)} \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{w - w_2^{(i)}}{\sigma_w^{(i)}} \right)^2} \theta(-w), \end{aligned} \quad (34)$$

where δ is the Dirac-delta function.

The first term in this formula shows that the ℓ_1 regularizer eliminates some of the assets from the portfolio by setting their weight to zero. The density of these assets, n_0 , is given by

$$n_0 = \frac{1}{N} \sum_i \left(\Phi \left(\frac{w_2^{(i)}}{\sigma_w^{(i)}} \right) - \Phi \left(\frac{w_1^{(i)}}{\sigma_w^{(i)}} \right) \right). \quad (35)$$

The two sums are made up of truncated Gaussians, the first sum corresponding to the weight distribution

of positive (long) positions, and the second to negative (short) ones. We see then that the series of discrete, sharp spikes in (15) is broadened by sample fluctuations, and in addition to the positive weights, also negative ones appear. Equation (34) reveals the meaning of the symbols introduced in (30), (31) and (32): $w_1^{(i)}$ and $w_2^{(i)}$ are the centers of the estimated positive resp. negative weight distributions of asset i , and $\sigma_w^{(i)}$ is the width of these distributions.

Note how the distribution of optimal weights has been obtained directly from our formalism, without having to go through the calculation of the estimated covariance matrix.

The order parameter q_0 will be of central importance for us. In [25] we showed that

$$q_0 \frac{1}{N} \sum_i 1/\sigma_i^2 = \tilde{q}_0 \quad (36)$$

is proportional to the *out-of-sample estimate* of the variance $\sum_{ij} w_i^{\text{est}} C_{ij}^{\text{true}} w_j^{\text{est}}$ as:

$$\tilde{q}_0 = \frac{\sum_{ij} w_i^{\text{est}} C_{ij}^{\text{true}} w_j^{\text{est}}}{\sum_{ij} w_i^{\text{true}} C_{ij}^{\text{true}} w_j^{\text{true}}}, \quad (37)$$

where C_{ij}^{true} is the true covariance matrix, w_i^{true} the corresponding optimal portfolio weights, and w_i^{est} is the optimal weights corresponding to the estimated covariance matrix. The denominator in (37) serves just to normalize \tilde{q}_0 . From the definition it is clear that $\tilde{q}_0 \geq 1$ and that

$$\sqrt{\tilde{q}_0} - 1 \quad (38)$$

is the *relative estimation error*.

3.1 Solution for complete information: $r \rightarrow 0$

The limit $r \rightarrow 0$ corresponds to $T \gg N$. This means we have much more data than the dimension, so in this limit we have to recover the results of Section 2.

From (30), (31), and (32) we see that $w_1^{(i)}$ and $w_2^{(i)}$ are of order $\mathcal{O}(1)$, while $\sigma_w^{(i)}$ vanishes. (In the small r limit λ and q_0 will be seen shortly to be of $\mathcal{O}(1)$, while Δ of $\mathcal{O}(r)$.)

Then, in the limit $r \rightarrow 0$ the arguments of the Ψ functions in (24) go to $+\infty$ and $-\infty$, respectively. For large x , $\Psi(x) \sim x$, and $\Psi(-x)$ is exponentially small, so (24) yields

$$\frac{1}{\sqrt{q_0 r}} \approx \frac{1}{N} \sum_i \frac{1}{\sigma_i} \frac{w_1^{(i)}}{\sigma_w^{(i)}},$$

which, by (30) and (32), leads to

$$\lambda = \frac{2}{\frac{1}{N} \sum_i 1/\sigma_i^2} + \eta_1, \quad (39)$$

in accordance with (11).

We anticipated that in the small r limit $\Delta \approx r$. Indeed, as $\lim_{x \rightarrow \infty} \Phi(x) = 1$ and $\lim_{x \rightarrow -\infty} \Phi(x) = 0$, (25) immediately gives $\Delta \approx r$, for $r \rightarrow 0$.

Finally, from (26) we obtain q_0 by noting that, for large x , $W(x) \sim x^2/2$ and $W(-x)$ is exponentially small:

$$q_0 = \frac{1}{\frac{1}{N} \sum_i 1/\sigma_i^2}, \quad \text{for } r \rightarrow 0. \quad (40)$$

Equation (40) then implies that for $r \rightarrow 0$, $\tilde{q}_0 \rightarrow 1$, which means that the relative estimation error vanishes, a natural result in the limit $T/N \rightarrow \infty$.

The value of the objective function at the stationary point is obtained by substituting the above results into equation (33):

$$f = \frac{1}{\frac{1}{N} \sum_i 1/\sigma_i^2} + \eta_1, \quad (41)$$

in agreement with (14).

Let us turn to the distribution of weights now. As the argument of the Φ s in (35) go to infinity for $r \rightarrow 0$, the Φ s themselves go to 1, so n_0 vanishes in this limit.

From (30) and (31) we see that for $r \rightarrow 0$ both set of weights $w_1^{(i)}$ and $w_2^{(i)}$ tend to

$$w_{1,2}^{(i)} \rightarrow \frac{1}{\sigma_i^2} \frac{1}{\frac{1}{N} \sum_j 1/\sigma_j^2}, \quad (42)$$

which is the same as the optimal weights found in (12).

In the same limit the standard deviations given in (32) vanish, so the Gaussians in (34) go over into Dirac delta functions. Since in the third term in (34) the delta spikes are multiplied by $\theta(-w)$, they do not contribute, so the distribution of weights in the $r \rightarrow 0$ limit becomes

$$p(w) = \sum_i \delta(w - w_i^*), \quad (43)$$

where w_i^* is the true optimal weights given in (12). We see then that in the limit $r \rightarrow 0$ our results derived via the replica method perfectly coincide with the results found in Section 2, thereby providing an important consistency check.

3.2 Including a riskless asset

If one of the assets, say the first, is riskless, $\sigma_1 \rightarrow 0$, then it must take on the full weight of the portfolio. (Remember that we have no constraint on the expected return of the portfolio, and are looking for the global minimum of the risk functional. Therefore, if there is a riskless asset in the portfolio, the total wealth must be invested in this asset.) Let us see how our equations lead to such a result.

As we will see later, above $r = 1$ zero modes (eigenvectors belonging to the proliferating zero eigenvalues of the estimated covariance matrix) appear in the system, and they start competing with the riskless asset. We will study these zero modes later; in the present subsection we restrict the discussion to the range $r < 1$, to avoid

the complications related to the zero modes. In order to ensure that our formulae remain meaningful, we assume that the product $N\sigma_1^2$ stays finite as $\sigma_1 \rightarrow 0$ and $N \rightarrow \infty$.

Now, if σ_1 is much smaller than the other variances, in (26) a single term dominates, and using the asymptotic behavior of $W(x) \sim x^2/2$, $x \rightarrow \infty$, we find

$$\frac{1}{N} \frac{(\lambda - \eta_1)^2 (1 + \Delta)^2}{4\sigma_1^2 q_0} = 1. \quad (44)$$

Similarly, from (24) we get

$$\frac{(\lambda - \eta_1)(1 + \Delta)}{2N\sigma_1^2} = 1. \quad (45)$$

Equations (44) and (45) imply

$$q_0 = N\sigma_1^2. \quad (46)$$

Then by (36) the quantity \tilde{q}_0 given in (37) is

$$\tilde{q}_0 = q_0 \frac{1}{N} \sum_i \frac{1}{\sigma_i^2} \approx q_0 \frac{1}{N\sigma_1^2} = 1. \quad (47)$$

As stated in (38), $\sqrt{\tilde{q}_0} - 1$ is the relative estimation error, so (47) means that the portfolio concentrated on the single riskless asset $i = 1$ is error free – an obvious result.

From (30), the weight $w_1^{(1)}$ is

$$w_1^{(1)} = \frac{(\lambda - \eta_1)(1 + \Delta)}{2\sigma_1^2}, \quad (48)$$

which, by (45) leads to

$$w_1^{(1)} = N, \quad (49)$$

so the riskless asset carries the total weight, indeed.

The only other weight that could compete with this is $w_2^{(1)}$, but it is positive and is multiplied by $\theta(-w)$ in the weight distribution, so it does not contribute, while all other weights are negligible in the $\sigma_1 \rightarrow 0$ limit.

Although according to (49) the riskless asset carries all the weight in the limit $\sigma_1 \rightarrow 0$, some small fluctuations still remain. The standard deviation given in (32) works out to be

$$\sigma_w^{(1)} = \sqrt{Nr}, \quad (50)$$

corresponding to Gaussian fluctuations about the average (49).

The above consideration is easily extended to the case when a few assets are much more stable than the others, in that their true volatilities are much smaller than those of the rest. Then this group of low volatility assets will take on almost all of the total weight N and share it among themselves in inverse proportion to their true variances, while the rest will receive only a small fraction of the weight. This distinction is encoded in the probability distribution of returns, and would show up in the

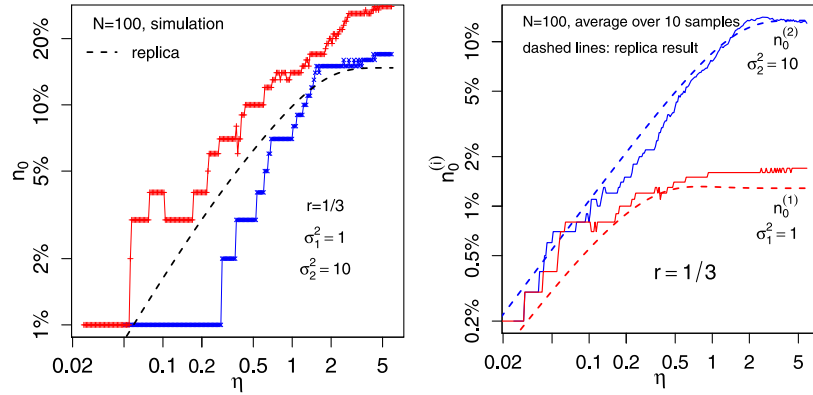


Fig. 1. Elimination of weights with increasing regularization parameter η . Left: the proportion of zero weights n_0 as function of η for two different single samples (blue and red) of a portfolio with the same composition of 100 assets. Note that the step-like functions are more or less following the trend of the theoretical curve (which has been derived in the large N limit and shown in the figure by a black dashed line), but the fluctuations for $N = 100$ are still large. Right: the step-like curves have been measured by averaging over 10 sample portfolios with the same composition: half of the 100 assets with $\sigma_1^2 = 1$, the other half with $\sigma_2^2 = 10$. The red dashed line shows the replica theoretic contribution of the $\sigma_1^2 = 1$ assets to the density of zero weights, the blue dashed line shows the same from the $\sigma_2^2 = 10$ assets. The contribution of the higher volatility component (shown in blue) is larger than that of the lower volatility one (in red): the regularizer eliminates the higher volatility assets with higher probability. Note that averaging over just 10 samples has substantially reduced the fluctuations.

estimated weights even in the absence of the regularizer. The regularizer's role becomes important when there is no such strong separation between the assets in the portfolio, and this is the case we are turning to now.

3.3 Elimination of assets by lasso

Before proceeding, we wish to emphasize again that we are calculating averages over the random samples, rather than trying to infer the behavior of the whole ensemble from studying a single sample. The difference is perhaps the most clearly seen in the case of the distribution of optimal portfolio weights. The lasso is known to eliminate some of the variables (setting their weights to zero). For a given sample with a given ratio $r = N/T$ this happens step-wise, i.e. as we increase the strength of the regularizer η (setting $\eta_1 = \eta_2 = \eta$ for simplicity) first one, then two, three, etc. weights will be rendered zero, in descending order of the corresponding variances. In contrast, the averaging over the samples in our formalism results in a density n_0 of zero weights that increases continuously with η and, according to (35), receives contributions from each asset i .

In Figure 1, we show numerical results for a two-variance portfolio and compare them to the results of the replica calculation. The numerical model is constructed from $N = 100$ assets, each having $T = 300$ data points ($r = 1/3$) drawn from a normal distribution. The variance of the returns is set to be $\sigma_2^2 = 1$ for half of the assets, while the other half has $\sigma_1^2 = 10$. As expected, the ℓ_1 regularizer mostly eliminates the weights associated with the higher variance group: in the right hand side figure $n_0^{(1)}$ indicates the proportion of the eliminated weights associated with the higher variance, while $n_0^{(2)}$ is the contribution of the lower variance assets. To indicate the size of fluctuations for a single portfolio, in the left figure the results for two different samples are shown,

compared to the replica result. From these figures one can form an idea how measurements performed on individual samples compare with the sample averages (at $r = 1/3$).

Let us now see what our theory has to say about the probability of the elimination of an asset depending on its variance. In line with what is suggested by the above measurement, one expects that more volatile assets will be removed with larger probability than the less volatile ones, that is the contribution to n_0 from asset i will be larger than that from asset j if $\sigma_i > \sigma_j$. Thus, we have to show that

$$\Phi\left(\frac{w_2^{(i)}}{\sigma_w^{(i)}}\right) - \Phi\left(\frac{w_1^{(i)}}{\sigma_w^{(i)}}\right) > \Phi\left(\frac{w_2^{(j)}}{\sigma_w^{(j)}}\right) - \Phi\left(\frac{w_1^{(j)}}{\sigma_w^{(j)}}\right). \quad (51)$$

If we introduce the notations

$$\frac{w_2^{(i)}}{\sigma_w^{(i)}} = z_i, \quad \frac{w_2^{(j)}}{\sigma_w^{(j)}} = z_j, \quad \frac{w_1^{(i)}}{\sigma_w^{(i)}} = y_i, \quad \frac{w_1^{(j)}}{\sigma_w^{(j)}} = y_j \quad (52)$$

then from (30)–(32) we see that

$$\frac{z_j}{z_i} = \frac{y_j}{y_i} = \frac{\sigma_i}{\sigma_j} = a > 1, \quad (53)$$

and

$$\frac{y_i}{z_i} = \frac{y_j}{z_j} = \frac{\lambda - \eta_1}{\lambda + \eta_2} = b < 1. \quad (54)$$

(The constant a is obviously positive, and it follows from (9) and (33) that $\lambda \geq \eta_1$, so b is non-negative.)

If we call $z_i = z$, the other three variables are simply proportional to it: $y_i = bz$, $z_j = az$, $y_j = abz$.

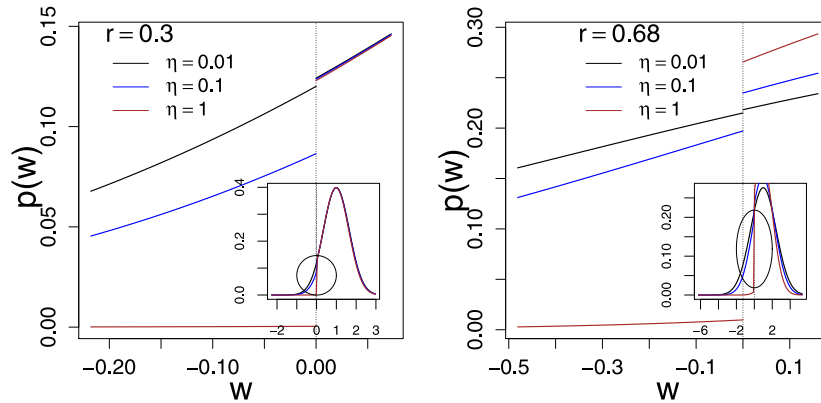


Fig. 2. The distribution of estimated weights with the ratio $r = 0.3$ resp. $r = 0.68$ and for different values of the regularizer's strength η when all the true standard deviations are the same, $\sigma_i = 1$ for all i , and $\eta_1 = \eta_2 = \eta$. Increasing η tends to suppress the negative weights. The vertical dotted line at the origin is meant to represent the Dirac-delta contribution of the zero weights.

The inequality (51) can then be written as

$$\Phi(z) - \Phi(bz) > \Phi(az) - \Phi(abz).$$

The definition of Φ , (27), then leads to

$$\int_{bz}^z dt e^{-t^2/2} > \int_{abz}^{az} e^{-t^2/2}, \quad a > 1, \quad (55)$$

so $f(z) = \int_{bz}^z dt e^{-t^2/2}$ must be a decreasing function of z . But $\frac{df}{dz} = e^{-z^2/2} - e^{-b^2 z^2/2} < 0$, indeed, because $b < 1$. Thus, we have shown that more volatile assets are eliminated from the portfolio by ℓ_1 with higher probability.

3.4 Resolution of portfolio weights

Turning now to the distribution of non-zero weights, we see from (34) that the discrete spikes in (43) split into two and get broadened by averaging over the samples. Figure 2 is an illustration of $p(w)$ in the special case when all the standard deviations σ_i are the same, $\sigma_i = 1$ for all i and $\eta_1 = \eta_2 = \eta$.

As we can see, with increasing r the Gaussians making up the distribution of weights become broader and broader, and the original sharp structure of $p(w)$ becomes washed away.

The question arises how small r must be in order to make it possible to resolve the structure of the weight distribution of a portfolio consisting of, say, just two classes of assets, $N/2$ assets with volatilities σ_i and $N/2$ with σ_j . A glance at Figure 3 shows that this is possible as long as the distance between the centers of the two Gaussians is larger than the mean of their standard deviations. From Figure 3 it is also clear that it is sufficient to consider the positive weights side of the distributions, so the requirement for resolvability is

$$\frac{(\lambda - \eta_1)(1 + \Delta)}{2\sigma_j^2} - \frac{(\lambda - \eta_1)(1 + \Delta)}{2\sigma_i^2} > \frac{1}{2} \left(\frac{\sqrt{q_0 r}}{\sigma_j} + \frac{\sqrt{q_0 r}}{\sigma_i} \right), \quad (56)$$

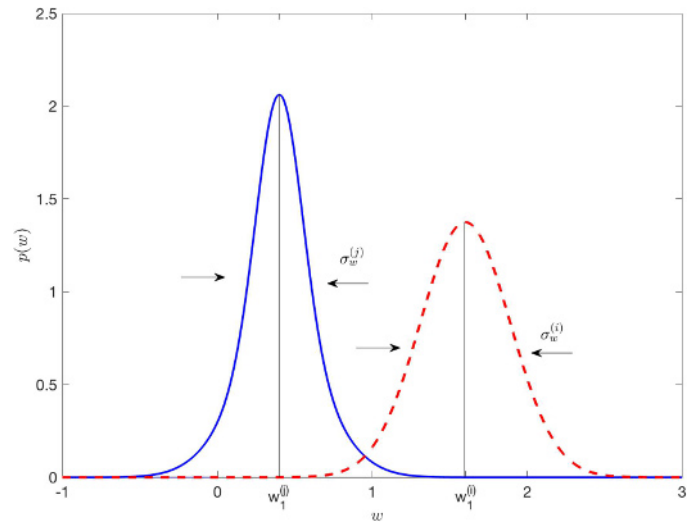


Fig. 3. The illustration of the resolution of two different assets.

that is

$$\frac{(\lambda - \eta_1)(1 + \Delta)}{\sqrt{q_0 r}} \left(\frac{1}{\sigma_j} - \frac{1}{\sigma_i} \right) > 1, \quad (57)$$

where we have assumed $\sigma_j < \sigma_i$.

When we have a large number of observations, i.e. $r \ll 1$, $\lambda - \eta_1$, q_0 , and Δ can be replaced by their $r = 0$ values, as given in (39), (40), and $\Delta = 0$, respectively. Then the resolvability of the two peaks will only depend on r and the two volatilities, and the criterion of resolvability becomes

$$\sqrt{\frac{r}{8}} < \frac{(\sigma_i - \sigma_j)}{\sqrt{\sigma_i^2 + \sigma_j^2}}, \quad (58)$$

where we have substituted σ_i for half of the assets and σ_j for the other half. It is then clear that for small r 's the inequality (57) is easily satisfied for σ 's sufficiently

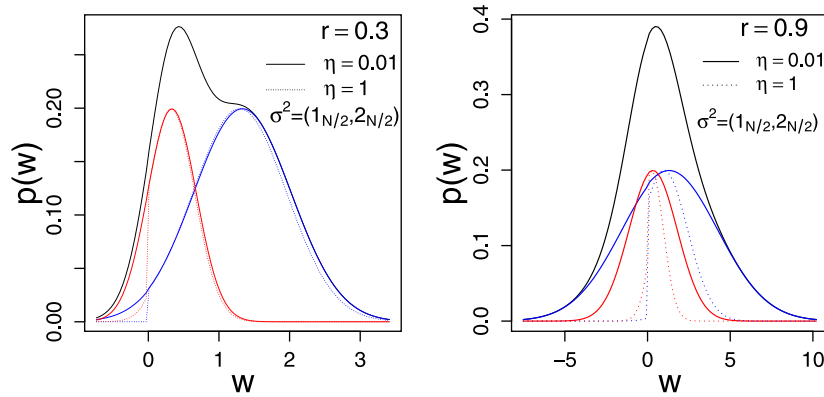


Fig. 4. Resolution of different assets for $r = 0.3$, resp. $r = 0.9$ for two different values of the regularization parameter η . The plot refers to the case where half of the assets have variance $\sigma_i = 1$, while the other half $\sigma_i = 2$, and $\eta \equiv \eta_1 = \eta_2$.

far apart. However, as will be seen shortly, with increasing r the coefficient $(\lambda - \eta_1)(1 + \Delta)$ on the left of (57) decreases rapidly, and the inequality gets violated: sample fluctuations will wash the structure away.

When one tries to estimate the portfolio weights from a single sample of empirical data, one is effectively picking the weights from the multimodal distribution $p(w)$ (like the distribution in Fig. 4, but with many more peaks). If the peaks are well separated and narrow, the estimates so obtained will be closed to the true weights, but this assumes small values of r , that is a large number of observations T . If T is not very large compared to N , the distribution of weights will lose its discrete structure, and the estimated weights may have very little to do with their true values. Figure 4 shows how much the discrete structure is already lost for $r = 0.3$, while for $r = 0.9$ there is no way to resolve the structure.

3.5 Results in the high-dimensional regime

In this subsection, we present results for the range of N and T values where their ratio is neither very small nor very close to $r = 2$. While at the two extremes it is easy to get analytic results by hand, in the intermediate r range one has to solve the first order conditions by the help of a computer. The results will be displayed below in a few figures. For comparison, the results for $\eta_1 = \eta_2 = 0$ (no regularization) and $\eta_1 = 0, \eta_2 \rightarrow \infty$ (no short positions allowed) are also shown. When the full regularizer is applied we set $\eta_1 = \eta_2 = \eta$, for simplicity. Also, since we have already displayed the results that depend on the heterogeneity of the portfolio (dominance of the riskless asset, preferential elimination of the large volatility items and the condition for the resolvability of nearby volatilities), we can henceforth set $\sigma_i = \sigma = 1$ for all i , for simplicity again.

Without regularization the optimization of variance does not have a meaningful solution beyond $r = 1$ where the first zero eigenvalues of the covariance matrix appear. Then q_0 and Δ diverge in the limit $r \rightarrow 1 - 0$, while λ and the in-sample estimate for the objective function vanish at $r = 1$. In the absence of regularization, the density n_0 of zero weights is identically zero.

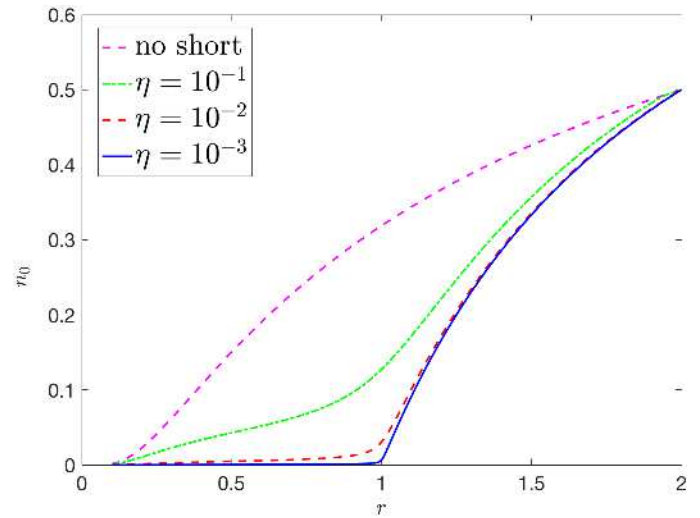


Fig. 5. The fraction n_0 of zero weights as function of r .

Regularization extends the region where the optimization can be carried out, from $0 \leq r < 1$ to $0 \leq r < 2$. We can see from Figure 5 that for small values of the coefficient η of the regularizer n_0 is very small for $r < 1$, but starts increasing fast above $r = 1$, ultimately going to $1/2$. Note that n_0 can be directly measured by numerical simulations; the agreement between the replica calculation and numerical simulation has already been shown in Figure 1.

There is a simple relationship between the density n_0 of zero weights and the order parameter Δ . From equations (25) and (35) one can see that

$$\Delta = \frac{r(1 - n_0)}{1 - r(1 - n_0)}. \quad (59)$$

The rapid growth of n_0 above $r = 1$ translates into a strong increase of Δ . (Without regularization Δ would diverge at $r = 1$.) With the regularizer on and n_0 going to $1/2$ as $r \rightarrow 2 - 0$, Δ ultimately diverges at $r = 2$. Equation (59) can serve as a recipe for the numerical determination of Δ through n_0 .

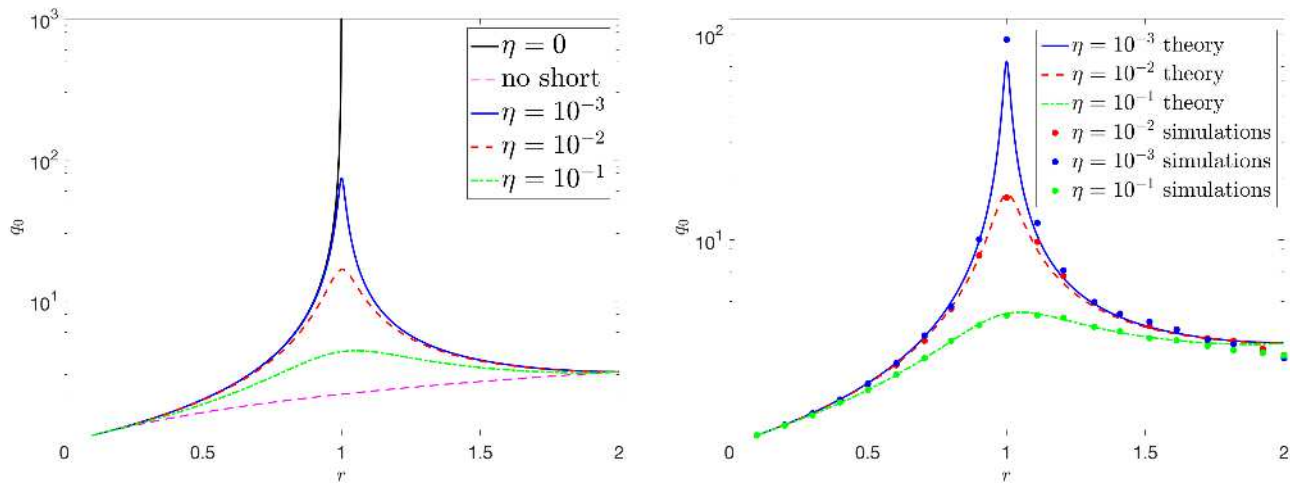


Fig. 6. Left panel: the behavior of the order parameter q_0 (proportional to the out-of-sample estimate for the variance, and also to the relative estimation error) as function of the ratio $r = N/T$ for different values of the coefficient η ($= \eta_1 = \eta_2$) of the regularizer. For small values of η , q_0 exhibits a sharp maximum (blue curve) around $r = 1$ where it would diverge without the regularizer. For larger η the maximum is less pronounced (dashed red curve), and for the largest value of $\eta = 0.1$ (continuous green curve) hardly any structure is noticeable around $r = 1$. Right panel: comparison with numerical simulations for different values of η and $N = 50$. The agreement between the analytic formula and numerical simulations is already good for a system of size $N = 50$.

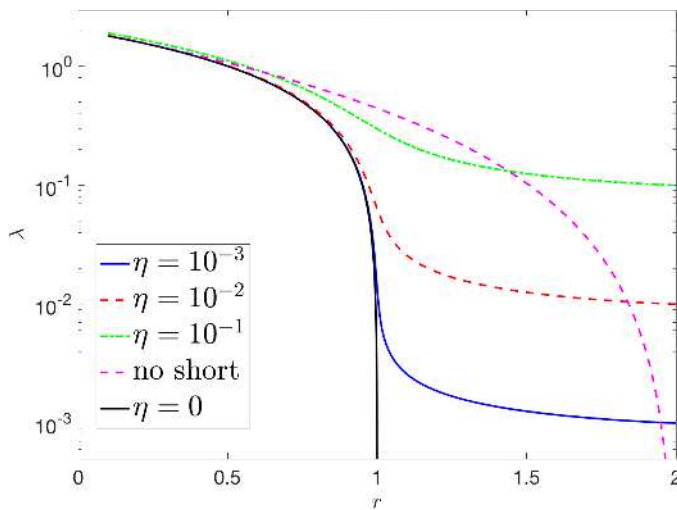


Fig. 7. The order parameter λ as function of r . Note the logarithmic scale on the vertical axis.

Figure 6 shows the behavior of the order parameter q_0 related to the estimation error and out-of-sample estimate for the objective function; q_0 is a quantity that can be obtained directly from simulations, the analytical and numerical results are compared in Figure 6 for various values of η . Without the regularizer q_0 would diverge at $r = 1$, similarly to Δ . As a vestige of this, for small values of the regularizer's coefficient η , q_0 shows a strong “resonance” around $r = 1$, but remains finite, and decreases above $r = 1$ to a finite limit. For larger η 's the resonance is suppressed, in particular, in the no-short-selling limit ($\eta_2 \rightarrow \infty$) q_0 is monotonically increasing over the entire interval $0 \leq r < 2$.

Finally, the in-sample estimator for the objective function f can be obtained from (33) through calculating λ from the stationarity conditions. The results for λ are exhibited in Figure 7. We shall see shortly that f goes to η_1 as $r \rightarrow 2 - 0$, implying that the variance vanishes in this limit.

3.6 Contour maps of estimation error

In order to assess the performance of regularization, we have to construct the contour lines of the estimation error. For simplicity we consider here a uniform portfolio with all the true variances $\sigma_i^2 = 1$, and for a first orientation let the left hand side slope η_2 of the regularizer go to infinity and keep the right hand side slope η_1 finite. The advantage of such an arrangement is that it excludes all the negative weights: $w_2^{(i)}$ defined in (31) goes to infinity, and $\Psi\left(-\frac{w_2^{(i)}}{\sigma_w^{(i)}}\right)$, $\Phi\left(-\frac{w_2^{(i)}}{\sigma_w^{(i)}}\right)$ and $W\left(\frac{-w_2^{(i)}}{\sigma_w^{(i)}}\right)$ all vanish in equations (24), (25) and (26). This leads to the much simplified set of equations:

$$\frac{1}{\sqrt{q_0 r}} = \Psi\left(\frac{(\lambda - \eta_1)(1 + \Delta)}{2\sqrt{q_0 r}}\right) \quad (60)$$

$$\Delta = \frac{r\Phi\left(\frac{(\lambda - \eta_1)(1 + \Delta)}{2\sqrt{q_0 r}}\right)}{1 - r\Phi\left(\frac{(\lambda - \eta_1)(1 + \Delta)}{2\sqrt{q_0 r}}\right)} \quad (61)$$

$$\frac{1}{2r} = W\left(\frac{(\lambda - \eta_1)(1 + \Delta)}{2\sqrt{q_0 r}}\right). \quad (62)$$

Applying the identity $W(x) = \frac{x}{2}\Psi(x) + \frac{1}{2}\Phi(x)$ in the last equation and using the previous two, after some simple manipulations one is led to the result that the arguments

of the functions Ψ , Φ and W above are equal to $\sqrt{\frac{\lambda - \eta_1}{2r}}$. Then the equations themselves become

$$\frac{1}{\sqrt{q_0 r}} = \Psi \left(\sqrt{\frac{\lambda - \eta_1}{2r}} \right) \quad (63)$$

$$\Delta = \frac{r\Phi \left(\sqrt{\frac{\lambda - \eta_1}{2r}} \right)}{1 - r\Phi \left(\sqrt{\frac{\lambda - \eta_1}{2r}} \right)} \quad (64)$$

$$\frac{1}{2r} = W \left(\sqrt{\frac{\lambda - \eta_1}{2r}} \right). \quad (65)$$

The last equation gives the solution for λ as

$$\sqrt{\frac{\lambda - \eta_1}{2r}} = W^{(-1)} \left(\frac{1}{2r} \right), \quad (66)$$

where $W^{(-1)}$ is the inverse of W . With r increasing λ is decreasing and goes to η_1 for $r \rightarrow 2$. As we have seen earlier, λ cannot be smaller than η_1 , so the square root remains real, and r cannot grow beyond 2. Substituting (66) into (64) and (65), respectively, we get the other two order parameters as functions of r . At first it may seem surprising that they depend only on r and do not depend on η_1 at all. (A little reflection shows that this is due to the combined effect of the exclusion of short positions and the budget constraint.) In particular, the order parameter q_0 , which determines the out-of-sample estimator for the objective function and the estimation error, works out to be

$$q_0 = \frac{1}{r} \frac{1}{\Psi^2 \left(W^{(-1)} \left(\frac{1}{2r} \right) \right)}. \quad (67)$$

This is independent of η_1 , but, of course, not independent of the regularization. Without regularization (and a very strong one at that; remember that we let $\eta_2 \rightarrow \infty$ at the beginning of this subsection) we would have $q_0 = \frac{1}{1-r}$ which is blowing up at $r = 1$, whereas (67) smoothly increases from 1 to π as r goes from zero to 2. Because q_0 is independent of η_1 , if we constructed the contour lines of q_0 , i.e. the lines of fixed q_0 on the $r - \eta_1$ plane, we would get a series of horizontal lines stacked above each other. (The above solution taken at $\eta_1 = 0$ is the optimization of the variance with a no-short constraint that we studied in [23].) When η_2 is finite we have to resort to a computer to construct the contour lines of q_0 . Now we set $\eta_1 = \eta_2 = \eta$, that is we consider a symmetric regularizer. The resulting q_0 contour lines are depicted in Figure 8a (This figure contains the same information as Fig. 6: the difference is that there q_0 was shown as a function of r , with the value of η as the parameter of the curves, while in here we are showing the constant q_0 lines on the $r - \eta$ plane, with q_0 as the parameter.)

We recognize the nearly horizontal contour lines immediately: in the lower regions of the figure (below $r = 0.3$, say) the lines of fixed q_0 are nearly independent of the

strength of the regularizer. As we go higher, the effect of the regularizer starts to be felt more and more. The estimation error $(\sqrt{q_0} - 1)$ on the first five contour lines, from bottom up, is 5%, 10%, 20%, 30%, and 40%, respectively. These lines are nearly horizontal, which means that if we have enough data the strength of the regularizer hardly matters at all, the error would be almost the same even for $\eta = 0$. The first line where we can see a definite increase of r with η is the one corresponding to the relative estimation error 0.4. Beyond this point the regularizer is taking over and the estimation error for a large enough η is determined more by the regularizer than the size of the sample. We see then that either we have a sufficient amount of data and then the regularizer does not play a very important role, or it does, but by then the error is so large as to make the whole optimization pointless. In the higher regions of the contour map, the optimization is completely determined by the regularizer. The highest contour line corresponds to $q_0 = \pi$ with r hitting its critical value of 2. For q_0 increasing further r must decrease (see Fig. 6). With q_0 going to infinity the contour lines shrink to the point $\eta = 0$, $r = 1$, corresponding to the singularity of the unregularized problem.

3.7 The critical behavior at $r = 2$

Let us start the analysis of the critical point with equation (26) and consider the general case where η_1 is different from η_2 ; we will see that η_2 does not appear in the results around $r = 2$. From equations (26), (30), (31) and (32) it is clear that the limiting behavior of the various quantities depends on λ and Δ , because q_0 remains finite here. The order parameter Δ diverges for $r \rightarrow 2 - 0$, but we will verify later that $\lambda - \eta_1$ goes to zero so fast that the product $(\lambda - \eta_1)(1 + \Delta)$ still vanishes at $r = 2$. At the same time $(\lambda + \eta_2)(1 + \Delta)$ diverges, therefore $W \left(\frac{-w_2^{(i)}}{\sigma_w^{(i)}} \right)$ vanishes, while according to

$$W(x) = \frac{1}{4} + \frac{x}{\sqrt{2\pi}} + \dots, \quad x \rightarrow 0 \quad (68)$$

the terms with $w_1^{(i)}$ become

$$W \left(\frac{w_1^{(i)}}{\sigma_w^{(i)}} \right) = \frac{1}{4} + \frac{1}{\sqrt{2\pi}} \frac{(\lambda - \eta_1)(1 + \Delta)}{2\sigma_i \sqrt{q_0 r}} + \dots \quad (69)$$

Then equation (26) becomes

$$\frac{1}{2r} = \frac{1}{4} + \frac{(\lambda - \eta_1)r(1 + \Delta)}{\sqrt{2\pi q_0 r}} \frac{1}{N} \sum_i \frac{1}{\sigma_i}, \quad (70)$$

or, to leading order in $\epsilon = 2 - r$,

$$\frac{\epsilon}{2} = \frac{(\lambda - \eta_1)\Delta}{\sqrt{\pi q_0}} \frac{1}{N} \sum_i \frac{1}{\sigma_i}, \quad (71)$$

which shows that the product $(\lambda - \eta_1)\Delta$ vanishes like $\sim \epsilon$ indeed.

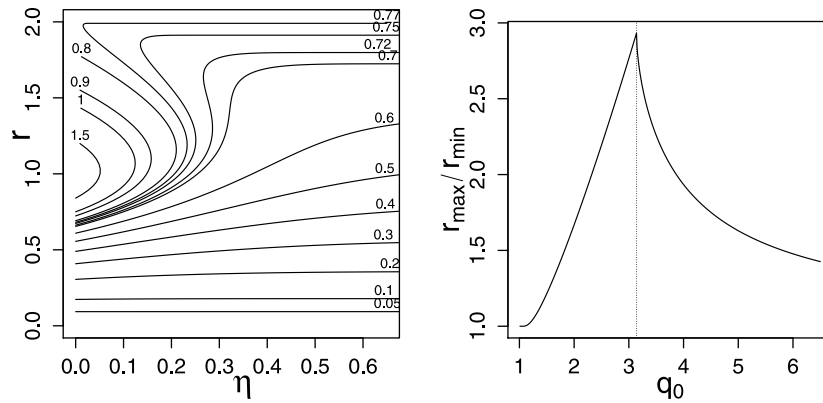


Fig. 8. (a) Contour plots of estimation error $\sqrt{q_0} - 1$ for ℓ_1 regularization with $\eta_1 = \eta_2 = \eta$. There is a critical value of q_0 at π , below which solution exists for any η . For low values of q_0 the result is almost insensitive to regularization. (b) Maximal improvement obtained by using regularization as a function of q_0 .

Similarly, from equation (24) with $\Psi(0) = \frac{1}{\sqrt{2\pi}}$ we get near $r = 2$

$$q_0 = \frac{\pi}{\left(\frac{1}{N} \sum_i \frac{1}{\sigma_i}\right)^2}, \quad r \rightarrow 2 - 0. \quad (72)$$

Accordingly, the $r \rightarrow 2 - 0$ limit of the relative estimation error given in (36) is

$$\tilde{q}_0 = q_0 \frac{1}{N} \sum_i 1/\sigma_i^2 = \pi \frac{\frac{1}{N} \sum_i 1/\sigma_i^2}{\left(\frac{1}{N} \sum_i \frac{1}{\sigma_i}\right)^2}, \quad (73)$$

where the expression multiplying π is, by force of the Cauchy inequality, larger or equal to 1 for any distribution of the true volatilities σ_i , therefore \tilde{q}_0 is larger than 1 for any distribution of the volatilities σ_i , as it should.

The asymptotic behavior of the order parameter Δ in (25) can be worked out similarly to obtain

$$\Delta = \frac{4}{2 - r}, \quad (74)$$

where use has been made of $\Phi(x) = \frac{1}{2} + \frac{x}{\sqrt{2\pi}} + \dots$, for x small.

Going back to (71) and using (72) and (74) we find

$$\lambda - \eta_1 = \frac{\pi}{8} \frac{\epsilon^2}{\left(\frac{1}{N} \sum_i \frac{1}{\sigma_i}\right)^2}, \quad (75)$$

vanishing quadratically for $r \rightarrow 2 - 0$. For the density of the zero weights we find

$$n_0 = \frac{1}{2} \quad (76)$$

in the same limit.

Turning to the distribution of weights, we see that $w_1^{(i)} \rightarrow 0$ for all i , so, in addition to the δ -peak at the

origin, all the positive weights collapse to zero, but with a finite standard deviation

$$\sigma_w^{(i)} = \frac{1}{\sigma_i} \frac{\sqrt{2\pi}}{\frac{1}{N} \sum_i \frac{1}{\sigma_i}}. \quad (77)$$

As for the weights $w_2^{(i)}$, they all go to infinity, so the corresponding contributions to (34) vanish exponentially.

Finally, the objective function can be obtained from (33). Here, the second term vanishes because of the divergence of Δ , while according to (75) the first term goes to η_1 , so

$$\lim_{r \rightarrow 2-0} f = \eta_1. \quad (78)$$

As we see, η_2 does not appear in any of the results near the critical point, but it is important to realize that its non-zero value ensures the vanishing of all the contributions with $w_2^{(i)}$.

What is happening at the transition at $r = 2$? To find the answer, we have to go back to the discussion below equation (6) where we found that the objective function $f \geq \eta_1$ and the equality only holds when the empirical variance vanishes and the optimal weight vector lies on the simplex. But then equation (78) implies that it is precisely this what is happening at the critical point. According to equation (6) the variance is the sum of T squares. This vanishes only if each T term vanishes separately. So, we need to find a weight vector that is pointing to the simplex and is orthogonal to the T random return vectors. This is exactly the same random geometry problem that we encountered in the case of the no-short-constrained optimization [23]. There we displayed a closed formula, valid for any N and T , for the probability of finding such a vector:

$$p(N, T) = \frac{1}{2^{N-1}} \sum_{k=T}^{N-1} \binom{N-1}{k}. \quad (79)$$

This formula depends only on the symmetry, and not on the concrete form of the return distribution, and as such, it is universal. For $N \leq T$ the probability of finding such a solution is zero. For N exceeding T the probability starts to increase, becomes $1/2$ at $N = 2T$ and goes to one as N increases further. If N and T go to infinity with their ratio $r = N/T$ held fixed, the function $p(N, T)$ goes over into a step function: the probability that the variance vanishes becomes zero for $0 < r < 2$ and 1 for $r > 2$. Thus, the critical point at $r = 2$ corresponds to a sudden transition between a situation where the variance is positive and one where it is zero with probability one, while the objective function becomes identically equal to η_1 , corresponding to a flat optimization landscape. This transition is similar to the large number of phase transitions in random high-dimensional geometry studied in [26] and [27].

4 Summary

We have considered the optimization of variance supplemented by a budget constraint and an asymmetric ℓ_1 regularizer. The present treatment includes as a special case the no-short-constrained portfolio optimization problem [23]. We have presented analytical results for the order parameter q_0 , directly related to the out-of-sample estimator of the objective function and the relative estimation error; for the in-sample estimator of the objective function; for the density of the assets eliminated from the portfolio by the ℓ_1 regularizer; and for the distribution of portfolio weights. We have studied the dependence of these quantities on the ratio r of the portfolio's dimension N to the sample size T , and on the strength of the regularizer. We have checked these analytic results by numerical simulations, and found general agreement. As the most conspicuous property of ℓ_1 is the step-like, one by one, elimination of the dimensions, we also run numerical experiments on single samples to reproduce this phenomenon. We have confirmed the appearance of the steps, and checked that the overall trend of the numerical results by and large follows the theoretical curve, which is remarkable, since the measurement was carried out on a single sample of finite size, whereas the theory is meant to work in the limit where both N and T go to infinity and the results are averaged over the whole ensemble of random samples. We have also seen that averaging over merely ten numerical curves is already enough to remove most of the fluctuations. We have repeatedly emphasized that the replica theory we applied in the analytic work is designed to average over infinitely many samples, and thus the results reflect the typical properties of the ensemble. Empirical work, in contrast, is usually dealing with a single sample, or a small number of samples, and tries to infer the properties of the ensemble from the information contained therein. Considering the rapid broadening with r of the Gaussians making up the distribution of weights, one can immediately see how misleading a small number of samples can be.

As portfolio optimization is just a simple representative example of quadratic optimization, our results have a message for these kind of optimization problems at large.

The extension of the interval where the optimization can be carried out, the maximal proportion of one half of dimensions eliminated by ℓ_1 and the “resonance” of the estimation error around the unregularized critical point at $r = 1$ are important findings – as is the disappointing performance ℓ_1 in the given context. The poor performance should not be a surprise, as in the given problem we were trying to rein in large fluctuations of a quadratic objective function by a regularizer which increases linearly. The phase transition taking place at $r = 2$ belongs to the large family of transitions in random geometrical problems studied in [26] and [27] where they were shown to be universal in the sense that the critical point is independent of the distribution of the data. As a manifestation of this universality, the critical value $r = 2$ does not depend on the Gaussian nature of the returns that we assumed here for the sake of easy application of the replica method.

We wish to point out that the present problem can be viewed as a statistical field theory model of a rather simple kind: an N -component ϕ^2 model in zero spatial dimensions with random masses. Because of the lack of a higher order term (such as a ϕ^4) in the model, there is no restoring force beyond the phase transition, instead we have a flat energy landscape. The budget constraint is a somewhat unusual feature of the model. If we replaced it by a spherical constraint, for example, we would immediately recognize the zero modes of the system as the Goldstone-modes. With the budget constraint combined with the ℓ_1 regularizer the sphere is replaced by the simplex, and the Goldstone-modes become the zero modes freely roaming about this invariant set, and the order parameter Δ blowing up at the $r = 2$ transition with an exponent -1 can be regarded as a kind of susceptibility associated with the transition. The extension of the concept of universality beyond the universality of the critical exponent also to the universality of the critical point itself is due to the purely geometric nature of the transition.

To conclude, we would like to call attention to the fact that the transition at $r = 2$ is very easy to overlook in empirical work. Upon approaching this critical point, the solution of the optimization problem as posed here would become unstable against “transverse” fluctuations which would leave the length of the weight vector approximately constant, but would result in large fluctuations in its direction. This corresponds to the weight vector freely roaming over the simplex. In finance terms, it would mean the optimal portfolio ending up with a different composition in each sample. It is clear that such a situation is undesirable (such a frequent rebalancing of the portfolio would be technically difficult and would result in high transaction costs), so the investor should keep well away from the point of instability. In numerical work, however, one may use, even inadvertently, some of the standard solvers that often contain a built in ℓ_2 regularizer without a clear warning about it. The presence of such “hidden” ℓ_2 regularizers in standard quadratic solvers has been pointed out in [23]. Such a regularizer will stabilize the solution and drive it toward the naive portfolio with all the weights equal. In a situation where the original problem is unstable even a very small ℓ_2 regularizer will suffice to do the job, thereby

creating the illusion that a stable solution can be obtained on the basis of a small number of observations.

In this paper, we have discussed the problem of optimizing a portfolio of independent assets, which is the worst possible case from the point of view of the sparsity of the underlying process that generates returns. Some of our findings, e.g. the value of the critical point or the number of eliminated dimensions, will certainly be modified when correlations between the assets are taken into account. Indeed, we expect the critical point to be shifted towards values of r larger than 2, and the maximum number of items that can be set to zero to increase beyond $N/2$ if the variance of returns is dominated by a few principal components. The study of the effect of correlations is however left for a later work.

Author contribution statement

All three authors contributed to the project design. F. Caccioli did most of the derivation of the replica free energy, I. Kondor did most of the solution of the saddle point equations, and F. Caccioli and G. Papp did the numerical calculations. I. Kondor wrote up most of the manuscript.

Open Access This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Appendix A: Derivation of the free energy with the replica method

As stated in the main text, (8), we have to find the optimum of the following objective function:

$$F = \frac{1}{T} \sum_{t=1}^T \left(\sum_i w_i x_{it} \right)^2 + \eta_1 \sum_i w_i \theta(w_i) - \eta_2 \sum_i w_i \theta(-w_i) - \lambda \left(\sum_{i=1}^N w_i - N \right) \quad (\text{A.1})$$

where the returns x_{it} are drawn from the joint probability density of independent Gaussian variables with zero mean and variance σ_i^2 .

Any optimization problem can be embedded into the formalism of statistical physics by regarding the objective function F as the “energy functional” of a fictitious system, introducing a fictitious inverse temperature γ , and integrating the Boltzmann factor $e^{-\gamma F}$ over the coordinates x_{it} in a given sample to get the “partition function” Z . The logarithm of the partition function $\ln Z$ is essentially a cumulant generating function from which all the quantities of interest can be obtained; in particular, the optimal weights can be found by minimizing the partition function over the weights in the “zero temperature” limit

$\gamma \rightarrow \infty$. The effectiveness of this procedure depends on the fact that we work in the limit of large N s where the distribution in the space of returns is extremely sharp around its maximum. The procedure just described gives us the optimal weights in a given sample of size T . However, if T is not much larger than the dimension N of the portfolio we are in the realm of high-dimensional statistics, where sample fluctuations are large, and optimizing our portfolio over a single sample can be very misleading. Therefore, in order to capture the typical properties, we have to average over the full ensemble of samples. This is analogous to averaging over the “quenched” random samples in the statistical physics of disordered systems [20], which explains why the methods developed in that theory can be successfully applied in the portfolio optimization context.

In order to average over the samples, we have to average the logarithm of the partition function which is a random variable fluctuating from sample to sample. Averaging the logarithm of a random variable is hard, while calculating the integer moments Z^n may be feasible. Now Z^n is just the partition function of n independent copies or replicas of the system (hence the name of the method). Assuming that we can analytically continue Z^n from the integers to real n s we can make the use of the identity

$$\langle (\ln Z)^n \rangle = \left\langle \frac{Z^n - 1}{n} \right\rangle, \quad (\text{A.2})$$

valid in the limit $n \rightarrow 0$.

Of course, the analytic continuation of a function from the integers to the reals is not necessarily unique. It is plausible, however, to assume that in the case of a convex objective function like that in (A.1), in the limit of large N all the replicas will go to the same minimum of $\ln Z$, and the simplest analytic continuation will do the job. Because we cannot provide a rigorous proof of this claim, we should regard the results of the replica calculation as heuristic. This is why we performed extensive numerical simulations to back up the analytic results in this paper. The general agreement we found is clear evidence of the correctness of the results. On the other hand, to deduce the nontrivial results from a purely numerical approach would have been obviously very hard if not impossible.

Let us now carry out the program sketched above. The replicated partition function is

See equation (A.3) next page

where $g(\vec{w}) = \eta_1 \sum_i w_i \theta(w_i) - \eta_2 \sum_i w_i \theta(-w_i)$ and, at an appropriate point, we will have to take the limits

$$\lim_{\gamma \rightarrow \infty} \lim_{n \rightarrow 0} \frac{1}{\gamma} Z_n(\vec{w}), \quad (\text{A.4})$$

where $\langle \dots \rangle$ represents an average over the probability distribution of returns.

The above partition function refers to a system of n replicas of the original system, and the index a is introduced to label different replicas, so that w_i^a represents the i th weight of the a th replica. Introducing an integral representation for the delta function and using the properties

$$Z_n(\vec{w}) = \left\langle \int_{-\infty}^{\infty} \prod_{i=1}^N \prod_{a=1}^n dw_i^a e^{-\gamma(\frac{1}{2} \sum_{i,j,t,a} w_i^a x_{it} x_{jt} w_j^a + \frac{T}{2} \sum_a g(\vec{w}^a)) + \frac{T}{2} \lambda(\frac{1}{N} \sum_i w_i - 1)} \right\rangle_{\vec{x}_t} \quad (\text{A.3})$$

$$Z_n(\vec{w}) = \left\langle \int_{-\infty}^{\infty} \prod_{i,a,t} dw_i^a d\phi_{at} d\lambda^a \exp \left[-\frac{1}{2} \sum_{a,t} \phi_{at}^2 + i\sqrt{\gamma} \sum_{i,t,a} \phi_{at}^a w_i^a x_{it} \right] \right. \\ \left. \times \exp \left[\frac{T}{2} \sum_a \lambda^a \left(\frac{1}{N} \sum_i w_i^a - 1 \right) - \frac{T\gamma}{2} \sum_a g(\vec{w}^a) \right] \right\rangle_{\vec{x}_t} \quad (\text{A.5})$$

$$Z_n(\vec{w}) = \int_{-\infty}^{\infty} \prod_{i,a,b,t} dw_i^a d\hat{Q}_{ab} d\phi_{at} d\lambda^a \exp \left[-\frac{1}{2} \sum_{a,t} \phi_{at}^2 - \frac{\gamma}{2} \sum_{a,b,t} \phi_{at} Q_{ab} \phi_{bt} \right] \\ \times \exp \left[\sum_{a,b} \hat{Q}_{ab} \left(N Q_{ab} - \sum_i \sigma_i^2 w_i^a w_i^b \right) + \frac{T}{2} \sum_a \lambda^a \left(\frac{1}{N} \sum_i w_i^a - 1 \right) - \frac{T\gamma}{2} \sum_a g(\vec{w}^a) \right] \quad (\text{A.6})$$

$$Z_n(\vec{w}) = \int_{-\infty}^{\infty} \prod_{i,a,b,t} dw_i^a d\hat{Q}_{ab} d\lambda^a \exp \left[-\frac{T}{2} \text{tr} \log (\delta_{ab} + \gamma Q_{ab}) \right] \\ \times \exp \left[\sum_{a,b} \hat{Q}_{ab} \left(N Q_{ab} - \sum_i \sigma_i^2 w_i^a w_i^b \right) + \frac{T}{2} \sum_a \lambda^a \left(\frac{1}{N} \sum_i w_i^a - 1 \right) - \frac{T\gamma}{2} \sum_a g(\vec{w}^a) \right] \quad (\text{A.7})$$

of Gaussian integrals the replicated partition function can be written as

See equation (A.5) above.

Averaging over the probability distributions of returns gives

See equation (A.6) above

where we have introduced the overlap matrix $Q_{ab} = \frac{1}{N} \sum_i \sigma_i^2 w_i^a w_i^b$ and the conjugate variables \hat{Q}_{ab} to enforce this relation.

We can now integrate over the variables ϕ_{at} to obtain

See equation (A.7) above.

It is at this point that we have to make the analytic continuation in the replica number n . In view of the permutation symmetry of the replicas and the convexity argument put forward earlier, we can choose the replica symmetric ansatz

$$Q_{ab} = \begin{cases} q_0 + \Delta, & a = b \\ q_0, & a \neq b \end{cases} \quad (\text{A.8})$$

$$\hat{Q}_{ab} = \begin{cases} \hat{q}_0 + \hat{\Delta}, & a = b \\ \hat{q}_0, & a \neq b. \end{cases} \quad (\text{A.9})$$

The analytic continuation will then consist in simply regarding n as a real variable. To leading order for small n we have

$$-\frac{T}{2} \text{tr} \log (\delta_{ab} + \gamma Q_{ab}) = -\frac{Tn}{2} \left[\log (1 + \gamma \Delta) + \frac{\gamma q_0}{1 + \gamma \Delta} \right] \quad (\text{A.10})$$

$$\sum_{a,b} \hat{Q}_{ab} Q_{ab} = Nn(\hat{q}_0 \Delta + q_0 \hat{\Delta} + \Delta \hat{\Delta}), \quad (\text{A.11})$$

while the \vec{w} -dependent part of the partition function can be written as

$$\int d\lambda^a d\hat{\Delta} d\hat{q}_0 \\ \times \exp \left[\frac{Nn}{2r} \left\langle \log \int dw e^{-2r\hat{\Delta}\sigma^2 w^2 + 2rwz\sigma\sqrt{-2\hat{q}_0} + \lambda w - g(\vec{w})} \right\rangle_{z\sigma} \right], \quad (\text{A.12})$$

where $\langle \cdots \rangle_{z,\sigma}$ denotes the average of an arbitrary function $h(z, \sigma)$ over the normal variable z and the distribution of asset variances:

$$\langle h(z, \sigma) \rangle_{z\sigma} = \int d\sigma \frac{1}{N} \sum_i \delta(\sigma - \sigma_i) \left(\int_{-\infty}^{\infty} \frac{dz}{\sqrt{2\pi}} h(z, \sigma) e^{-z^2/2} \right). \quad (\text{A.13})$$

If we now write the partition function as

$$Z_n = \int d\lambda dq_0 d\Delta d\hat{q}_0 d\hat{\Delta} e^{-\gamma n \frac{T}{2} F(\lambda, q_0, \Delta, \hat{q}_0, \hat{\Delta})}, \quad (\text{A.14})$$

we find for F/N

$$\begin{aligned} f(\lambda, q_0, \Delta, \hat{q}_0, \hat{\Delta}) &= \frac{1}{\gamma} \left[\log(1 + \gamma\Delta) + \frac{\gamma q_0}{1 + \gamma\Delta} \right] + \frac{\lambda}{\gamma} \\ &\quad - \frac{2r}{\gamma} (\hat{q}_0\Delta + q_0\hat{\Delta} + \Delta\hat{\Delta}) \\ &\quad - \frac{1}{\gamma} \left\langle \log \int dw e^{-2r\hat{\Delta}\sigma^2 w^2 + 2rwz\sigma\sqrt{-2\hat{q}_0} + \lambda w - g(\vec{w})} \right\rangle_{z\sigma}. \end{aligned}$$

Noting how the various quantities scale with the inverse temperature we can perform the change of variables $\Delta \rightarrow \Delta/\gamma$, $\hat{q}_0 \rightarrow \gamma^2 \hat{q}_0$, $\hat{\Delta} \rightarrow \gamma \hat{\Delta}$, $\lambda \rightarrow \gamma\lambda$ and taking the limit $\gamma \rightarrow \infty$ we finally have

$$\begin{aligned} f(\lambda, q_0, \Delta, \hat{q}_0, \hat{\Delta}) &= \frac{q_0}{(1 + \Delta)} - 2r\hat{q}_0\Delta - 2r\hat{\Delta}q_0 + \lambda \\ &\quad + \min_{\vec{w}} \left\langle V(\vec{w}) \right\rangle_{z\sigma}, \end{aligned} \quad (\text{A.15})$$

where

$$V = 2r\hat{\Delta}\sigma^2 w^2 - 2rwz\sigma\sqrt{-2\hat{q}_0} - \lambda w + \eta_1\theta(w) - \eta_2\theta(-w). \quad (\text{A.16})$$

This is the form of the objective function that we use in the main text. Its minimization is explained in Section 3.

References

1. J.D. Jobson, B. Korkie, Proc. Am. Stat. Assoc. (Bus. Econ. Stat.) **1**, 279 (1979)
2. P. Jorion, J. Financ. Quant. Anal. **21**, 279 (1986)
3. O. Ledoit, M. Wolf, J. Empir. Finance **10**, 603 (2003)
4. O. Ledoit, M. Wolf, J. Portfolio Manage. **31**, 110 (2004)
5. O. Ledoit, M. Wolf, J. Multivar. Anal. **88**, 365 (2004)
6. V. Golosnoy, Y. Okhrin, Eur. J. Finance **13**, 441 (2007)
7. T. Shinzato, J. Stat. Mech.: Theor. Exp. **2017**, 023301 (2017)
8. O. Ledoit, M. Wolf, Ann. Stat. **40**, 1024 (2012)
9. J. Bun, J.-P. Bouchaud, M. Potters, My beautiful laundrette: Cleaning correlation matrices for portfolio optimization, 2016, available at <https://www.researchgate.net/publication/302339055> (accessed: September 2017)
10. O. Ledoit, M. Wolf, Direct nonlinear shrinkage estimation of large-dimensional covariance matrices, University of Zurich, Department of Economics, 2017, Working paper No. 264, p. 46
11. R. Tibshirani, J. R. Stat. Soc. Ser. B (Methodological) **58**, 267 (1996)
12. R. Jagannathan, T. Ma, J. Finance **58**, 1651 (2003)
13. J. Brodie, I. Daubechies, C. De Mol, D. Giannone, I. Loris, Proc. Natl. Acad. Sci. **106**, 12267 (2009)
14. V. DeMiguel, L. Garlappi, F.J. Nogales, R. Uppal, Manage. Sci. **55**, 798 (2009)
15. D. Giomouridis, S. Paterlini, J. Financ. Res. **33**, 223 (2010)
16. M. Carrasco, N. Noumon, Optimal portfolio selection using regularization. University of Montreal, 2012, <http://www.unc.edu/maguilar/metrics/carrasco.pdf> (accessed: September 2017)
17. J. Fan, J. Zhang, K. Yu, J. Am. Stat. Assoc. **107**, 592 (2012)
18. Y.-M. Yen, T.-Y. Yen, Comput. Stat. Data Anal. **76**, 737 (2014)
19. B. Fastrich, S. Paterlini, P. Winker, Quant. Finance **14**, 2019 (2014)
20. M. Mézard, G. Parisi, M.A. Virasoro, in *Spin Glass Theory and Beyond*, World Scientific Lecture Notes in Physics (World Scientific, Singapore, 1987), Vol. 9
21. F. Caccioli, I. Kondor, M. Marsili, S. Still, Int. J. Theor. Appl. Finance **19**, 1650035 (2016)
22. T. Hastie, R. Tibshirani, J. Friedman, in *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, 2nd edn. Springer Series in Statistics (Springer, Berlin, 2008)
23. I. Kondor, G. Papp, F. Caccioli, J. Stat. Mech.: Theor. Exp. **2017**, 123402 (2017)
24. P. Bühlmann, S. Van De Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications* (Springer Science Plus Business Media, Berlin, Heidelberg, 2011)
25. I. Varga-Haszonits, F. Caccioli, I. Kondor, J. Stat. Mech.: Theor. Exp. **2016**, 123404 (2016)
26. D. Donoho, J. Tanner, Philos. Trans. R. Soc. A, Math. Phys. Eng. Sci. **367**, 4273 (2009)
27. D. Amelunxen, M. Lotz, M.B. McCoy, J.A. Tropp, Inform. Inference **3**, 224 (2013)